WILEY-VCH

molecular informatics

RESEARCH ARTICLE OPEN ACCESS

CoLiNN: A Tool for Fast Chemical Space Visualization of Combinatorial Libraries Without Enumeration

Regina Pikalyova | Tagir Akhmetshin | Dragos Horvath | Alexandre Varnek

Laboratoire de Chemoinformatique, University of Strasbourg, Strasbourg, France

Correspondence: Alexandre Varnek (varnek@unistra.fr)

Received: 3 September 2024 | **Revised:** 3 February 2025 | **Accepted:** 28 February 2025

Keywords: chemical space | combinatorial library | compound enumeration | DNA-encoded libraries (DELs) | GTM

ABSTRACT

Visualization of the combinatorial library chemical space provides a comprehensive overview of available compound classes, their diversity, and physicochemical property distribution - key factors in drug discovery. Typically, this visualization requires time- and resource-consuming compound enumeration, standardization, descriptor calculation, and dimensionality reduction. In this study, we present the Combinatorial Library Neural Network (CoLiNN) designed to predict the projection of compounds on a 2D chemical space map using only their building blocks and reaction information, thus eliminating the need for compound enumeration. Trained on 2.5 K virtual DNA-Encoded Libraries (DELs), CoLiNN demonstrated high predictive performance, accurately predicting the compound position on Generative Topographic Maps (GTMs). GTMs predicted by CoLiNN were found very similar to the maps built for enumerated structures. In the library comparison task, we compared the GTMs of DELs and the ChEMBL database. The similarity-based DELs/ChEMBL rankings obtained with “true” and CoLiNN predicted GTMs were consistent. Therefore, CoLiNN has the potential to become the go-to tool for combinatorial compound library design – it can explore the library design space more efficiently by skipping the compound enumeration.

1 | Introduction

Combinatorial chemistry allows to produce a vast number of structurally diverse molecules by simple and repetitive steps of covalent building block (BB) linkage. The expansion of combinatorial compound collections has allowed scientists to explore chemical spaces comprising billions of molecules. Thus, chemical suppliers such as OTAVA, WuXi LabNetwork, and Enamine offered collections exceeding 50 billion make-on-demand molecules, whereas proprietary chemical spaces of pharmaceutical companies have developed much larger collections of virtual molecules. GlaxoSmithKline′s GSK XXL space stands as the largest compound collection to date, containing approximately 10^{26} entries [1, 2].

An interesting example of combinatorial collection are DNA-Encoded Libraries (DEL) containing molecules synthesized from building blocks covalently attached to DNA tags [3–15]. DNA encoding allows to screen all compounds in a DEL simultaneously in a mixture against a biological target of interest, thus allowing to explore large regions of the chemical space all at once [16]. According to estimations of Goodnow [15], the upper bound of DEL is about 10^{12} compounds. Theoretical analysis of virtual combinatorial libraries before their actual synthesis and testing allows to identify potentially hit-enriched compound libraries before any experimental procedures are undertaken and, in such a way, significantly reduce the human and material costs [16].

Various methods, such as QSAR modelling [17], docking [17], and dimensionality reduction [18–20], can be used to analyze combinatorial libraries. However, despite their utility, these methods require an enumeration of compounds structure. When dealing with larger libraries or several large-sized collections, enumeration no longer seems feasible due to the immense processing resources required and 'big data' storage issues [15].

The analysis of vast combinatorial libraries without enumeration was first addressed in the works of Rarey et al. [21–25]. The authors exploited the fact that combinatorial libraries (CL) are composed of different building blocks combined according to specific reaction rules to efficiently analyze, search, and compare even non-enumerable combinatorial spaces. The main principle of these methods is to represent each molecular fragment by fingerprint representations that capture their chemical and topological features or encode a compound as a tree of its fragment features describing the ability of forming interactions. For example, the Ftrees [21] methodology enables fast query-based similarity searches in CL spaces based on feature trees. Tools like SpaceProp [25] and SpaceCompare [24] extend this approach: SpaceProp calculates property distributions without full enumeration, while SpaceCompare facilitates the comparison of different combinatorial chemical spaces using fragment fingerprint representations. Nevertheless, the visualization of the chemical space using dimensionality reduction methods is unattainable without first enumerating the compounds.

The first method for visualization of the chemical space without the need for compound enumeration was described in the work of Agrafiotis and Lobanov [26]. They developed a three-layer fully connected Multi-Layer Perceptron (MLP) trained to predict multidimensional scaling projections on a 2D map of combinatorial products using as input descriptors of their respective building blocks. This approach was tested on a two-building block combinatorial library based on reductive amination reaction with 90 K compounds. However, their method did not consider the reactions used to create the library. This shortcoming complicates the analysis of combinatorial libraries with multiple reaction types, requiring multiple reaction-specific models instead of a single unified model.

Here, we propose a Combinatorial Library Neural Network (CoLiNN) which, given the building blocks and reactions of the combinatorial library, predicts the projection of a product on a 2-dimensional latent space without compound enumeration. Unlike previously reported MLP model [26], CoLiNN creates both building blocks and reactions embeddings followed by assessment of compound projections using the *softmax* activation function.

As a dimensionality reduction method, we chose the Generative Topographic Mapping [27–33] (GTM), successfully used for the analysis of large chemical collections [27–33]. Unlike other popular methods like PCA, UMAP or t-SNE presenting a compound projection by two coordinates (X and Y), GTM provides a fuzzy projection characterized by a so-called "responsibility" vector rendering the association degree of the projected items to each of the nodes of a rectangular grid defining the map. This responsibility vector, which can be visualized as a density "blob" on the map, can be cumulated to describe coverage of chemical space by a library, and herewith enable a direct high-level library comparison approach [16, 17]. Fast prediction of these cumulated responsibility patterns of combinatorial libraries is paramount to establishing a procedure for library design powered by chemical space overlap considerations (find the best building block in order to achieve coverage of a defined "patch" of chemical space on the map). If CoLiNN is able to predict a responsibility vector (here, of 1681 dimensions), it can easily be rebuilt to output the classical (X,Y) latent coordinates of most dimensionality reduction methods [34–37]. This was not attempted here, because fuzzy responsibilities are already the tool of choice for chemical space overlap analysis.

Below we describe the CoLiNN architecture and related predictive models trained on several DNA-Encoded Libraries. Obtained results demonstrate efficiency of the developed tool to visualize (ultra)large combinatorial libraries using information about building blocks and chemical reactions, and avoiding structure enumeration step.

2 | Data

2.1 | DELs

In this work, we used DELs containing from 1 M to 7B compounds designed and enumerated in our previous study [18]. They resulted from combinatorial synthesis involving 2- or 3-BB libraries and 2 or 3 reaction steps. The standardization of enumerated molecules was done using the same procedure described below for BBs.

Two CoLiNN models were developed:

1. A local (library-specific) model trained on a small subset of compounds from one 80 M-sized DEL
2. A general chemistry-sensitive model trained on subsets from 388 DELs

2.2 | Building Blocks (BBs)

The set of commercially available BBs from eMolecules Inc. [13]. and Enamine [39] used for DEL enumeration in our previous study [18] was taken as input to the CoLiNN model. BBs were standardized using ChemAxon Standardizer according to the procedure implemented on the Virtual Screening Web Server of the Laboratory of Chemoinformatics at the University of Strasbourg. This process includes dearomatization and final aromatization (heterocycles like pyridone are not aromatized), dealkalization, conversion to canonical SMILES, removal of salts and mixtures, neutralization of all species, except nitrogen(IV), generation of the major tautomer according to ChemAxon. After standardization and duplicate removal 70 691 unique BBs were obtained. For each of them, a unique identifier starting from 0 was given. For training of the global CoLiNN model, only 388 DELs with unique reaction schemes were used that employ 64 869 BBs for their enumeration from the total BB set.

2.3 | DEL for Local CoLiNN Model

The local CoLiNN model was trained on compounds from DEL2568. It is a 3-BB DEL based on three reactions: aldehyde reductive amination, Migita thioether synthesis, and guanidinylation of amines. This DEL was fully enumerated by eDesigner and after removing duplicates had a size of 81 M compounds. For training, a random subset of 1 M compounds was used. Four CoLiNN models were trained on four training sets to select the minimum required training set size: 1) 1 M set, 2) 50 K random compounds taken from 1 M, 3) 25 K random compounds taken from 1 M, 4) 10 K random compounds taken from 1 M. The remaining 80 M compounds from this DEL (excluding the 1 M representative set) were used for testing.

2.4 | DELs for Global CoLiNN Model

For global CoLiNN training, we identified how many DELs are needed to get a diverse training set out of the space of 2473 DELs we generated originally. For this, all DELs were clustered according to their reaction schemes. Here reaction scheme stands for a particular sequence of two or three reactions in a defined order used for DEL enumeration. This resulted in 388 different clusters, i.e., unique reaction schemes not accounting for deprotection reactions. See the bar plot showing the number of DELs per reaction scheme in Figures S1 and S2 of the Supporting Information. Per each reaction scheme, one DEL was taken for training CoLiNN. Only 10 K compounds from each of these DELs were taken for training, giving rise to a 3 880 000 training set compounds.

2.5 | ChEMBL

ChEMBL database version 28 was taken here as a reference dataset to be used for comparison to DELs. ChEMBL28 was filtered according to the rules of DEL-likeness introduced in our previous work [19]. Standardization of ChEMBL molecules was done in the same way as for BBs described above. After duplicate removal, the size of the filtered ChEMBL was 1 605 370 compounds.

2.6 | Reactions

Reaction information was input for CoLiNN training. They were taken from the Supporting Information document of the eDesigner [40] enumeration tool. It contains numerical codes and names of reactions as encoded in eDesigner. In this study, we assigned each reaction a unique index ranging from 1–29 to be used for CoLiNN training. The reaction names and the correspondence between eDesigner numerical codes and indices used herein are given in Table S1 of the SI.

3 | Methods

3.1 | Generative Topographic Mapping (GTM)

Generative Topographic Mapping [29] (GTM) is an unsupervised method for dimensionality reduction based on manifold learning. The GTM algorithm consists in optimizing the manifold shape to fit the data represented in the multidimensional descriptor space. The optimization is done by tuning manifold hyperparameters, such as its flexibility and smoothness (determined by the number of Radial Basis Functions (RBFs) usually radially symmetric Gaussians, RBF width factor σ, and the spacing of RBFs), map size, and regularization coefficient. When the optimal form of the manifold is found, data points are projected to it with node-specific probabilities called responsibilities (r_{ik} – the responsibility of the molecule i to be projected to the node k). In such a way, a data point (molecule) can be projected to multiple nodes at the same time meaning it is associated with several chemotypes [33, 41]. Finally, a manifold is flattened back to its planar form giving an interpretable 2D map with projected on it input dataset compounds.

When multiple molecules of the chemical library are projected onto the GTM, a cumulative responsibility for each node k, see Equation (1) can be calculated, which is roughly equal to the number of compounds residing there. The cumulative responsibility values per node can be rendered using a colour code allowing to visualize the quantitative distribution of compounds across the chemical space giving rise to a density landscape. In this work, density landscapes are used to visualize the chemical space of the analysed DELs.

$$\mathbf{c_k} = \sum_{\mathbf{i}}^{\mathbf{N}} \mathbf{r_{ik}} \tag{1}$$

r_{ik} is the responsibility value of the molecule i in the node k

However, to capture library size-independent chemical space coverage, $\mathbf{c_k}$ should be normalized by the library size N, resulting in $\phi = (\phi_1, \ldots, \phi_k)$ vector to represent the entire chemical library, see Equation 2.

$$\phi_{\mathbf{k}} = \frac{\mathbf{c_k}}{\mathbf{N}} \tag{2}$$

Here, ϕ is calculated either from the predicted or GTM-derived responsibility vectors. To compare the GTM-derived ϕ and ϕ' predicted by CoLiNN of the same library a Tanimoto similarity coefficient (Tc) can be calculated, see Equation (3). Its value indicates how much the predicted and the "true" GTM-derived maps are similar to each other.

$$\boldsymbol{Tc}(\phi\ ,\phi'\ ) = \frac{\sum_k^K \phi_k \cdot \phi'_k}{\sum_k^K \phi_k^2 + \sum_k^K \phi'^2_k - \sum_k^K \phi_k \cdot \phi'_k} \tag{3}$$

K total number of nodes on the map

3.2 | Combinatorial Library Neural Network (CoLiNN)

3.2.1 | Inputs

A product molecule (DEL compound) was represented as a sequence of reactions and BB identifiers as shown in Figure 1.

In this work, the BB was represented as an undirected hydrogen-labelled graph [42] (HLG) where edge type or bond order (single, double, triple, aromatic) is taken into account using the number of hydrogens each atom is connected to. In more detail, the HLG feature vector is defined by atomic number, period, group, number of electrons plus atom′s charge, shell, an indication of whether an atom is in a ring, number of neighboring atoms, and the total number of hydrogens connected to each atom. The latter allows to account for bond order directly in the feature vector avoiding the use of relational GCN where each layer has r weight matrices, where r is the number of unique bond types. Instead, in our GCN, only one weight matrix per layer is used owing to the additional value in the feature vector. The features described herein were calculated using the EPAM Indigo toolkit [43] and graphs were created using PyTorch Geometric [44].

3.2.2 | Outputs

The target value used for training CoLiNN is a responsibility vector of the compound derived from the GTM as shown in Figure 1. The responsibility vector is composed of r_{ik} values that represent probabilities of a compound to be projected on to the map. It describes which nodes (k) of the map a molecule (i) is projected to and thus can be called a projection vector as well. The responsibility vectors used herein for DELs were all derived from the universal GTM that was trained on ChEMBL database [45] frame set and was used to visualize the space of DELs in our previous works [18–20].

3.3 | Architecture

CoLiNN is a Graph Convolutional Network that performs two tasks during training: (1) it creates and saves embeddings for Building Blocks (BBs), and (2) it creates and trains on molecule embeddings that are assembled from BB and reaction embedding vectors to be able to predict responsibility vectors of compounds (see Figure 2 and Figure 3). The steps of this architecture are as follows:

1. Embedding Calculation for Building Blocks (BBs): Each BB is assigned a unique ID, and its embedding vector is calculated as shown in Figure 2. First, atom feature vectors are linearly transformed into initial embedding vectors of dimension D. Then the latter as well as the adjacency matrix are passed through five Graph Convolution Network (GCN) layers. The mathematical transformation occurring in each GCN layer is described in Equation (4). The obtained atom embedding vectors in this way are then summed up to get a BB embedding vector and saved in a file.
2. Molecule representation: The molecule is represented by a sequence of reaction and BB indices. Then, saved BB embeddings are associated with corresponding indices. Reaction embedding vectors are obtained by converting the reaction indices into numerical vectors of dimension D.
3. Application of masks: A mask is applied to the BB embeddings to zero out those generated for padding BBs. For instance, in a 2-BB compound, the embedding for the third "padding" BB is zeroed out.

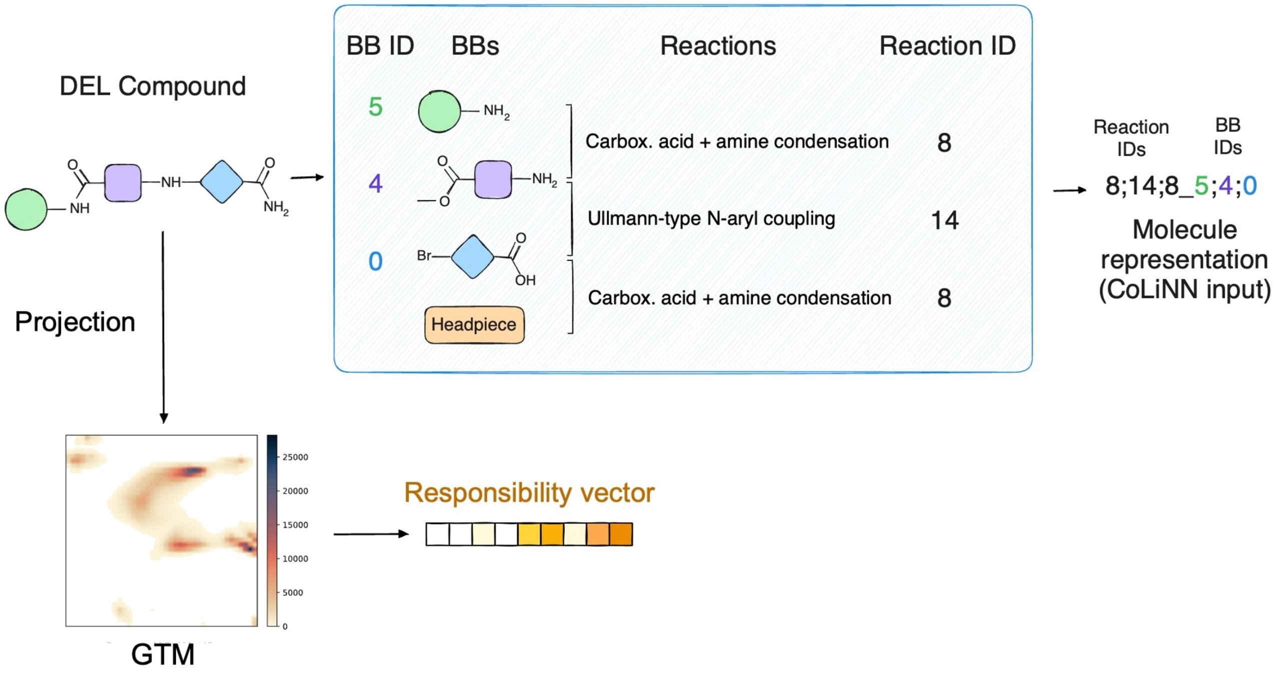


FIGURE 1 | Creation of the product molecule representation that is used as input for CoLiNN.

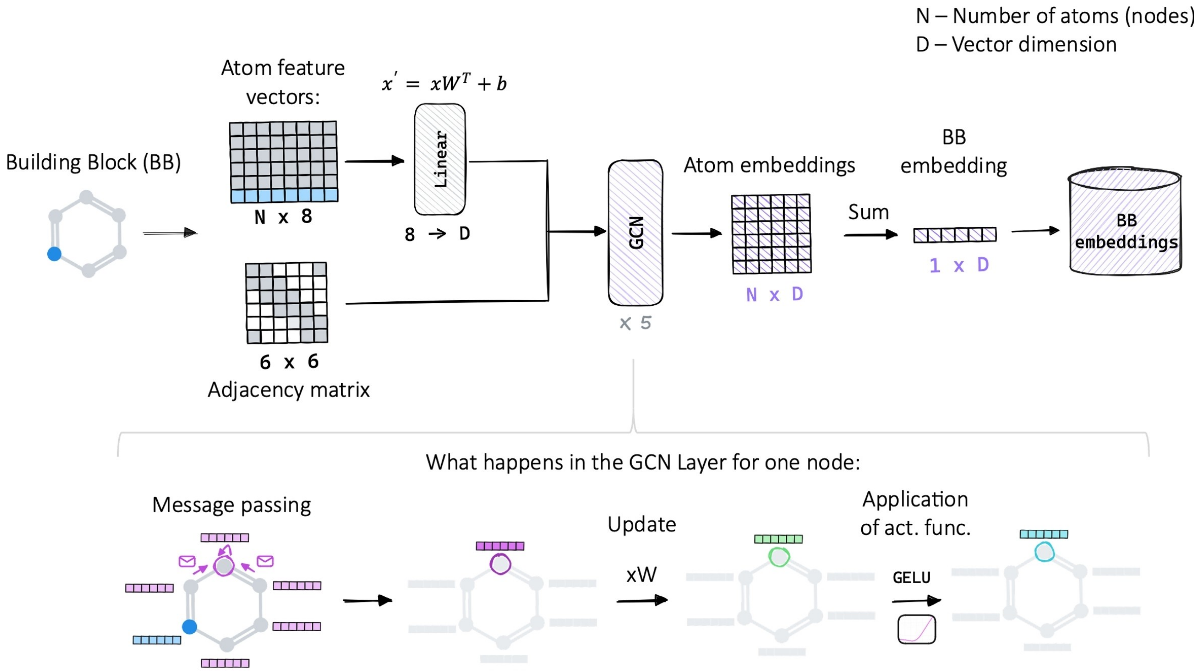


FIGURE 2 | Calculation of building block (BB) embedding vectors.

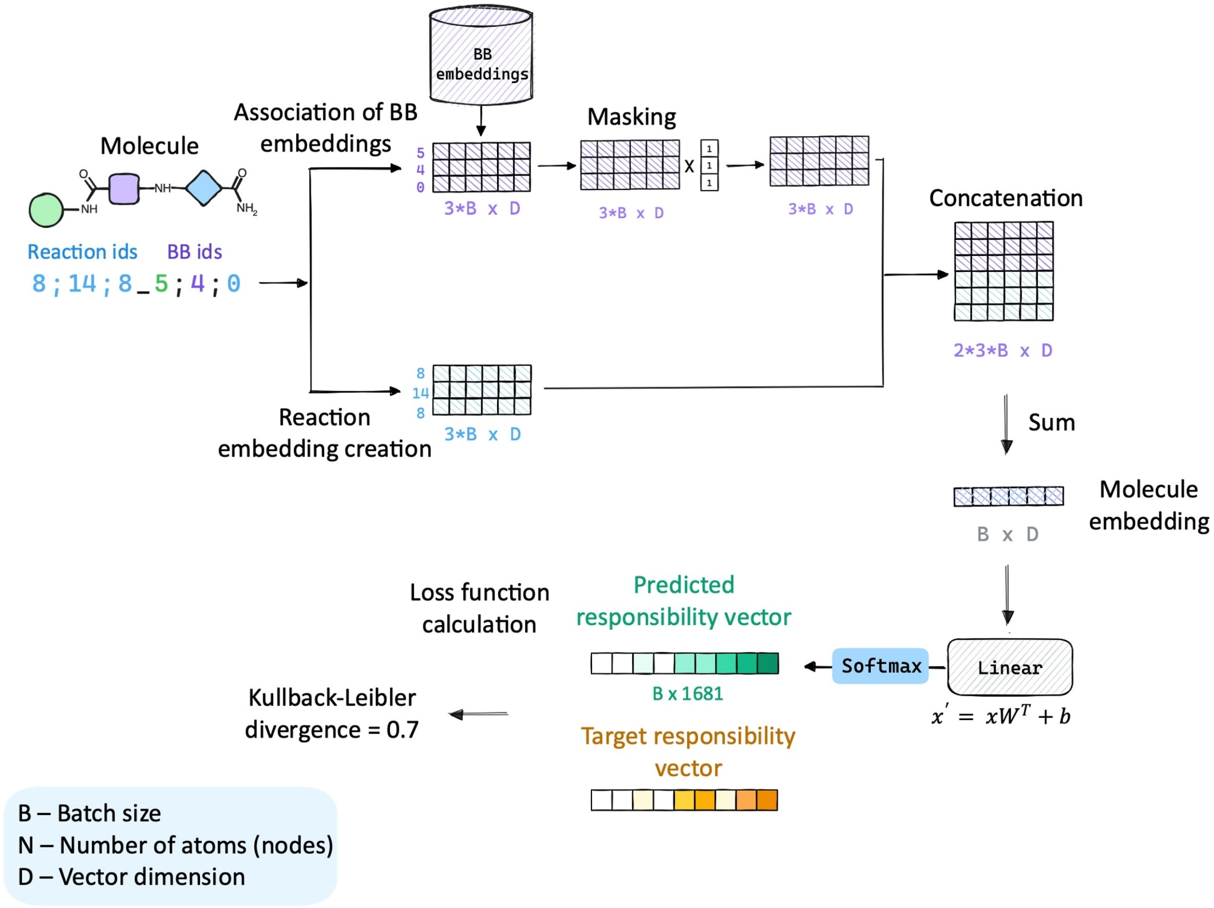


FIGURE 3 | Scheme explaining the training process of CoLiNN. Here, for simplicity, every step is shown for the batch size of one molecule.

4. Assembly of the full molecule vector: The full molecule vector is created by combining BB and reaction embeddings. They are concatenated to integrate information about both molecular composition and reactions used to

enumerate the compound. They are then summed up to form a vector representing the molecule.

5. Responsibility vector prediction: The molecule vector is linearly transformed and a softmax activation function is applied to give an output 1681-dimensional responsibility vector, which represents the compound projection on the GTM of size 41 × 41 nodes.
6. Loss function calculation: Kullback-Leibler divergence is calculated between the responsibility vector predicted by CoLiNN and the actual enumerated product projection; see Equation (5).

By following these steps, CoLiNN effectively encodes the complex structure and reaction information of molecules into a form suitable for responsibility prediction tasks without compound enumeration.

CoLiNN is coupled with 5 Graph Convolution Network layers where GELU [46] (Gaussian Linear Unit) is used as an activation function. Instead of ReLU used in the original implementation of GCN by Kipf and Welling [47] the GELU activation was chosen as it is smoother than ReLU and is differentiable at every point [48]. This helps to have an improved gradient flow during backpropagation [48, 49] and to reduce the number of dead neurons [48]. Mathematically, GCN operation can be described by the following equation:

$$\mathbf{H}^{(\mathbf{l}+1)} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{\mathbf{l}}\mathbf{W}^{\mathbf{l}}\right) \tag{4}$$

$A \in R^{n\times n}$	adjacency matrix,
$A_{i,j}$	1, if there is an edge between nodes i and j 0, otherwise
$I \in R^{n\times n}$	identity matrix
$\tilde{D} \in R^{n\times n}$	diagonal degree matrix of $\tilde{A}$
$H^l \in R^{n\times d}$	per-node feature vectors or node embeddings from the previous layer
$W^l \in R^{d\times w}$	weight matrix for layer l
σ	non-linear activation function, here Gaussian Error Linear Unit (GELU)
n	the number of nodes
d	the number of node features

3.4 | Loss Function

The responsibility vector predicted by CoLiNN was compared to the one obtained using the GTM algorithm using Kullback-Leibler divergence, see Equation (5). The latter measures how the inferred probability distribution diverges from the expected probability distribution.

$$\textbf{KL div}\,(\mathbf{R_i} \,\|\, \mathbf{R'_i}) = \sum_{\mathbf{k}=1}^{\mathbf{K}} \mathbf{r_{ik}} \cdot \mathbf{ln}\,\frac{\mathbf{r_{ik}}}{\mathbf{r'_{ik}}} \tag{5}$$

$\mathbf{R_i} = (\mathbf{r_{i1}}, \mathbf{r_{i2}}, \ldots, \mathbf{r_{ik}})$	true responsibility vector of the molecule i
$\mathbf{R'_i} = (\mathbf{r'_{i1}}, \mathbf{r'_{i2}}, \ldots, \mathbf{r'_{ik}})$	predicted responsibility vector of the molecule i

4 | Results and Discussion

4.1 | Local CoLiNN Model

The local CoLiNN model is specific to DEL2568. To select the minimum required training set size, four CoLiNN models were trained and tested using four subsets of sizes 10 K, 25 K, 50 K and 1 M DEL2568. This 1 M set was in all cases excluded from testing, performed over the remaining 80 M compounds from the full DEL. The training and validation KL div loss values for all local models are given in Table S2 in SI.

In Figure 4, the reference density landscape of the 80 M DEL generated using the GTM algorithm and those predicted by CoLiNN are shown. As the training set size increases from 10 K to 1 M, the similarity to the reference GTM-produced landscape also increases. However, if instead of comparing the maps visually, we look at the Tanimoto similarity coefficient between the GTM-generated maps and predicted maps Tc(ϕ,)ϕ' in Figure 5, a plateau of Tc is observed at 25 K already. This means that the minimum size required to train a robust model is 25 K compounds. This CoLiNN model is able to predict the characteristic GTM projection pattern of the 80 M-sized library in 2 hours instead of the 13 days taken by the standard procedure: eDesigner [40] enumeration, standardization, ISIDA fragment descriptor calculation, projection of compounds on the GTM.

4.2 | Global CoLiNN Model

The global model was trained on 10 K samples from 388 DELs employing unique reaction schemes (see Table S1 in SI). 50 epochs were sufficient to achieve convergence, which corresponds to a training duration of CoLiNN of about 13 h. Training and validation KL div loss values for different epochs are reported in Table 1.

The trained global CoLiNN was tested both on all compounds from 388 DELs not used for training (990 K compounds per DEL) and all the other 2089 DELs that did not participate in training at all. Albeit they did not serve for training, these latter share building blocks and reaction schemes with the 388 training DELs. Figure 6 shows the distribution of Tanimoto similarity values between the “true” and predicted ϕ vectors for all DELs. In the vast majority of cases, predictions are accurate

TABLE 1 | Values of the loss (KL divergence) at the beginning, halfway, and end of training CoLiNN on 388 DELs.

	Train loss	Validation loss
Epoch 1	4.66	2.49
Epoch 50	0.78	0.79
Epoch 100	0.73	0.77

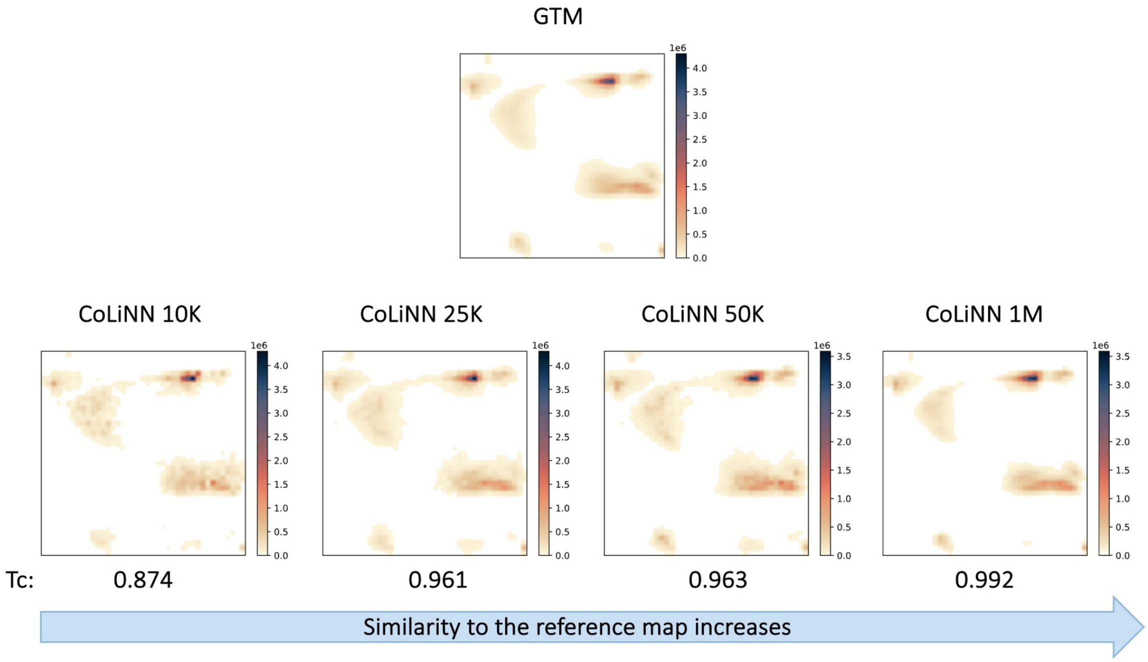


FIGURE 4 | Density landscapes of DEL2568 containing 80 M compounds, from left to right: landscape generated by GTM algorithm, predicted by CoLiNN trained on 10 K, 25 K, 50 K, and 1 M compounds.

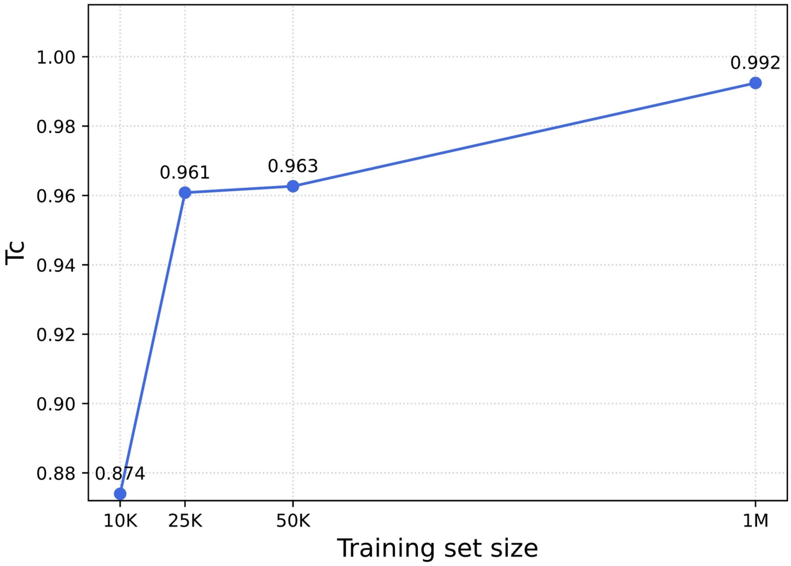


FIGURE 5 | Line plot showing the Tc(ϕ,)ϕ' for different CoLiNN models trained on 10 K, 25 K, 50 K, and 1 M compounds.

(for almost all train set and for 1600 of test set DELs ≈80% of all DELs) but a few notable failures occur. Examples of badly and perfectly predicted DEL maps (both from the test set) are given at the bottom of Figure 6. CoLiNN almost perfectly predicted the responsibility vectors for all compounds from 3BB DEL1953 with Tc(ϕ,ϕ' being 0.99. This is a 3BB library based on Schotten-Baumann coupling between amines and sulfonyl chlorides, 1,2,3-triazole synthesis and aldehyde reductive amination. For the DEL117 the predicted map is significantly dissimilar from the true one as reflected by the low Tc(ϕ,ϕ') of 0.07 – a few density peaks are in the wrong places on the landscape. It is a 2BB library based purely on reductive aminations. The badly predicted DELs with the lowest Tc, were analyzed, but no trend was found in terms of the number of BBs, reactions, or degree of inclusion of their BBs in the train set. More examples of predicted maps for internal set and external set DELs are given in Figure 7 and Figure 8. The Figure S3 in the SI provides additional analysis of the validity of CoLiNN predictions using KL divergence calculated between predicted and true responsibility vectors. It shows the distribution of the KL divergence values for 10 DELs out of 2089 and 10 DELs out of 388.

To see the performance of the global CoLiNN model applied to the visualization of an ultra-large compound library, it was tested for prediction of the map of 80 M-sized DEL2568. This

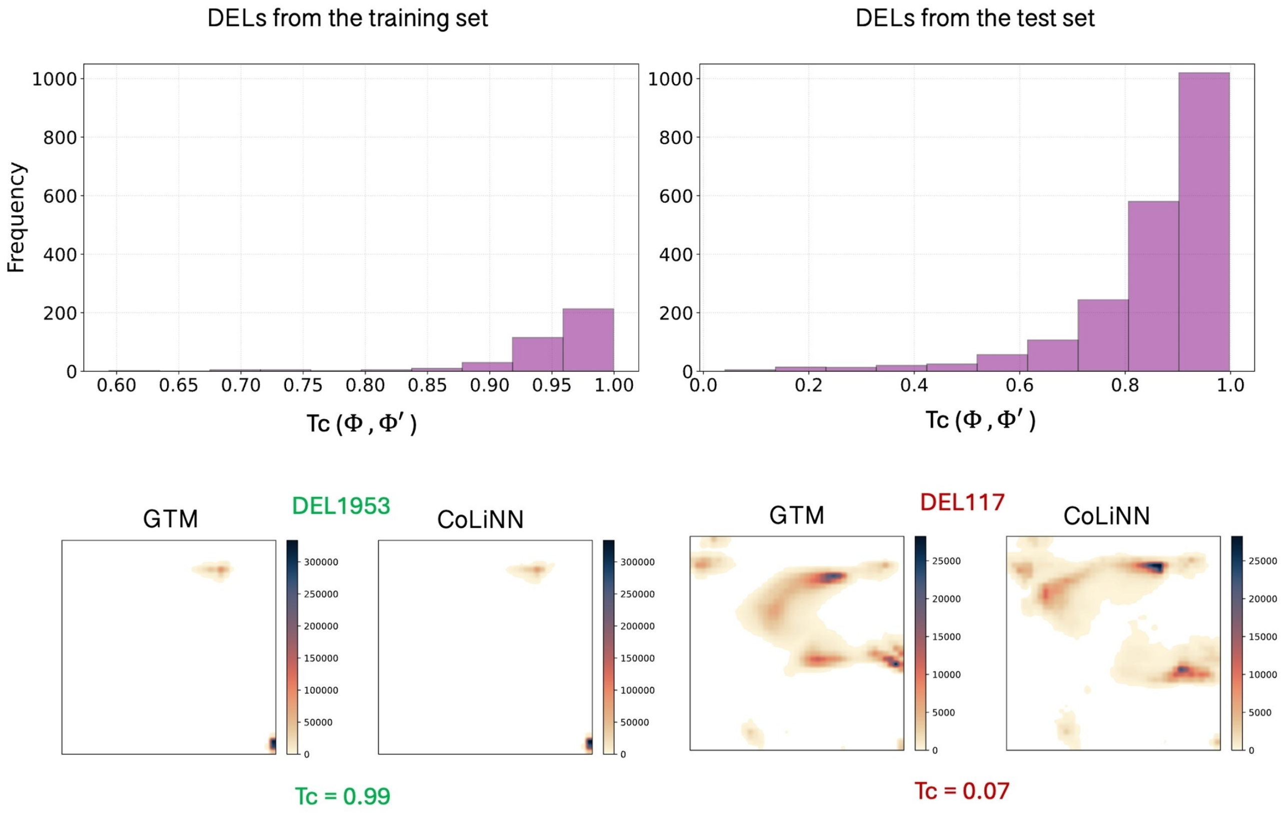


FIGURE 6 | Histograms showing the distribution of Tc(ϕ,)ϕ' for either training or test set DELs. Examples of badly and perfectly predicted DEL maps (from the test set) visualized as density landscapes are given below.

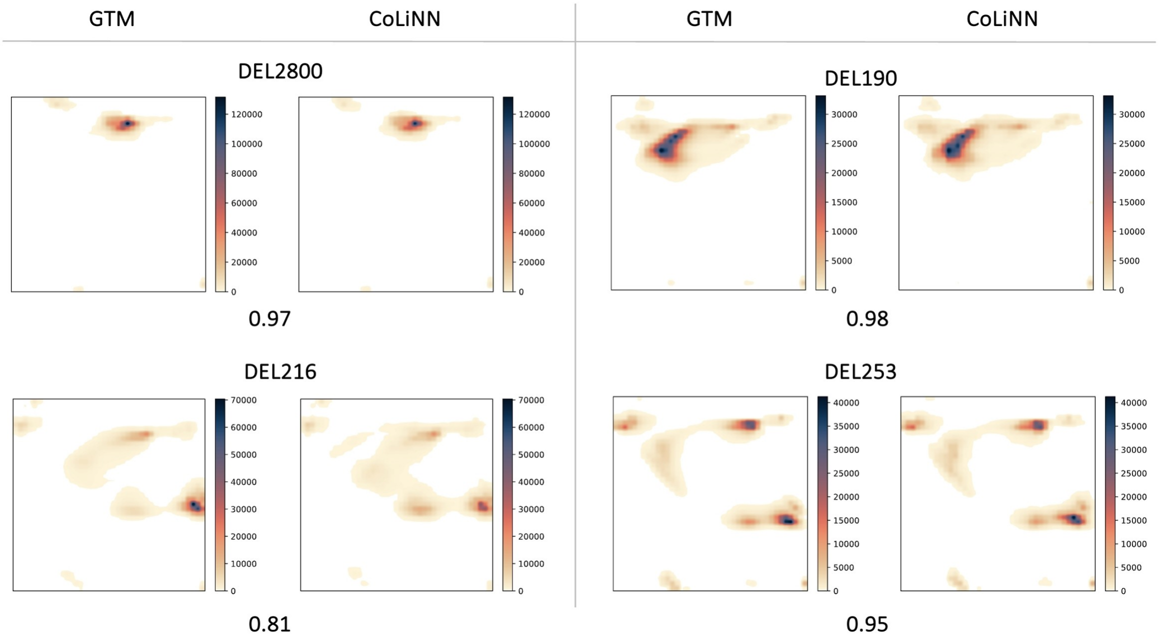


FIGURE 7 | Density landscapes of four DELs that participated in the training, herein only 990 K compounds absent from the training set were projected. Landscapes on the left are GTM-produced, the ones on the right were predicted by CoLiNN. DEL216 and DEL190 are 2BB DELs, all others are 3BB libraries. The values underneath the two maps correspond to the Tc similarity between them.

particular DEL was not a part of 388 DELs used for global CoLiNN training and validation. The predicted density landscape of this DEL2568, shown in Figure 9 mimics very well the true density landscape produced by the GTM algorithm. Tc between them is 0.91, showcasing the high accuracy of the global CoLiNN model for the prediction of even multimillion-sized combinatorial library maps.

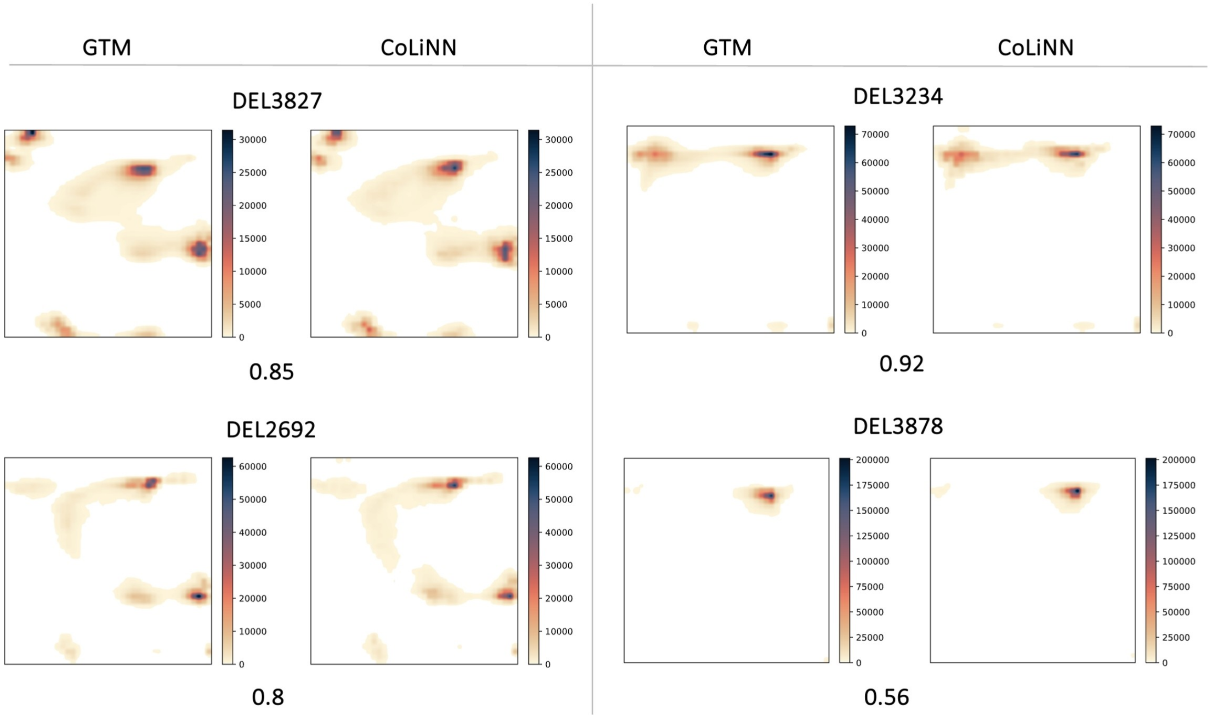


FIGURE 8 | Density landscapes of four external test set DELs that did not participate in CoLiNN training, herein all 1 M compounds per DEL are projected. Landscapes on the left are GTM-produced, the ones on the right were predicted by CoLiNN. All are 3BB libraries. The values underneath the two maps correspond to the Tc similarity between them.

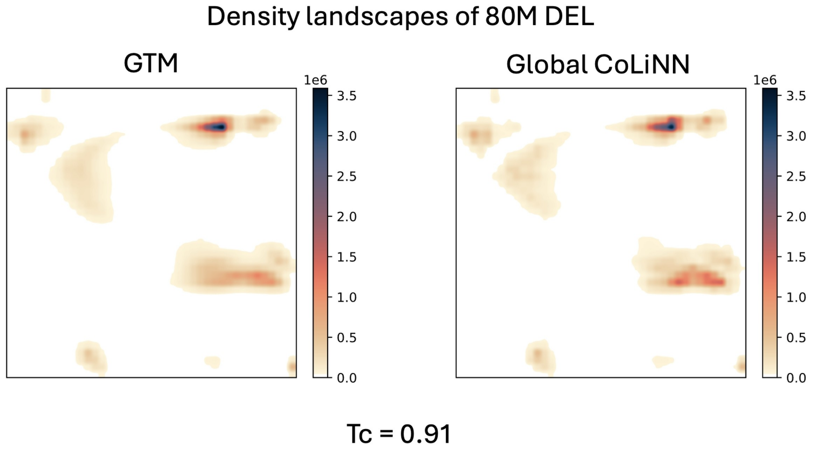
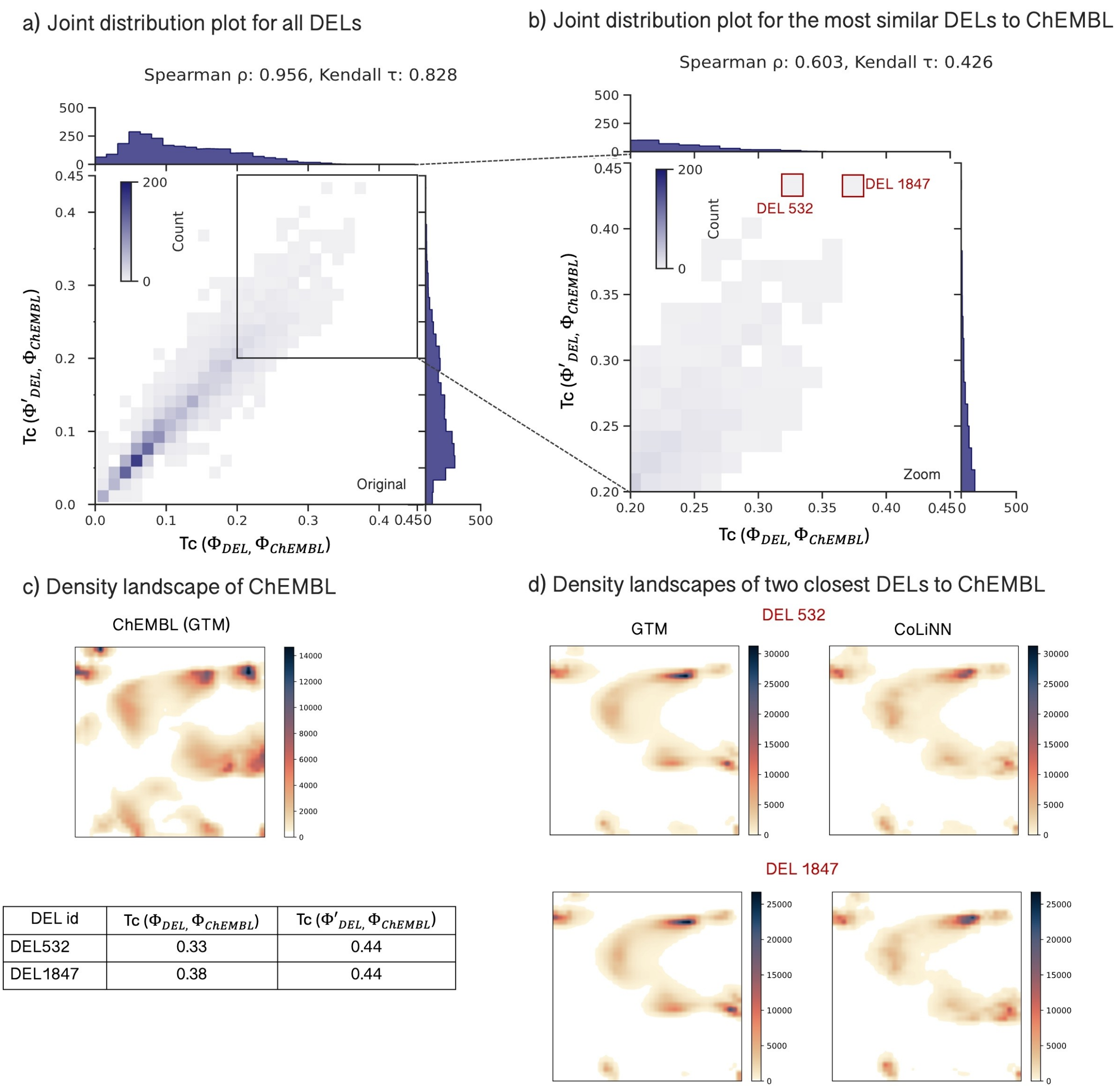


FIGURE 9 | Density landscapes of ultra-large DEL2568 containing 80 M compounds. On the left: density landscape calculated using GTM algorithm. On the right: density landscape predicted by Global CoLiNN.

Notwithstanding, the main idea behind CoLiNN is to predict correctly the "library chemical space motif" to be able to accelerate chemical library comparison. Hence, further on we will look at how well the ranking by similarity of DELs to ChEMBL (reference database) is preserved when using GTM-calculated or CoLiNN-predicted maps. For this, we calculated the Tc($\phi_{DEL}, \phi_{ChEMBL}$) based on true GTM responsibilities and Tc($\phi'_{DEL}, \phi_{ChEMBL}$) based on responsibilities predicted by CoLiNN, their distribution is given in Figure 10a. To compare the rankings of DELs according to Tc similarity to ChEMBL obtained either by GTM or CoLiNN, Spearman ρ and Kendall τ correlation coefficients were used. The two rankings correlate as reflected by high Spearman ρ =0.956 and Kendall τ=0.828. The correlation between Tc values is also visually apparent from the joint distribution plot in Figure 10a. However, what is more important here to look at is the closest DELs to ChEMBL with Tc=0.2–0.44. A zoomed joint distribution plot in Figure 10b shows that there are DELs for which the predicted similarity diverges from the true similarity. As an example, two such DELs (DEL532 and DEL1847) highlighted in red on the plot were analyzed. Their Tc values and density landscapes calculated using GTM or predicted by CoLiNN are given in Figure 10d. For comparison purposes, the density landscape for ChEMBL28 is given in Figure 10c. Their density landscapes clearly show that CoLiNN did not mispredict the library chemical space motif – the GTM-calculated map and predicted



DEL id	Tc ($\Phi_{DEL}, \Phi_{ChEMBL}$)	Tc ($\Phi'_{DEL}, \Phi_{ChEMBL}$)
DEL532	0.33	0.44
DEL1847	0.38	0.44

FIGURE 10 | a) Joint distribution plot showing the correlation between Tc(ϕDEL, ϕChEMBL) and Tc(ϕ'DEL, ϕChEMBL), as well as individual Tc distributions on the margins. The brightness represents the density of data points (counts). b) A zoomed version of the joint distribution plot showing the correlation between Tc(ϕDEL, ϕChEMBL) and Tc(ϕ'DEL, ϕChEMBL) for the closest DELs to ChEMBL, two DELs that show less correlation are highlighted in red. c) Density landscape of ChEMBL28 and a table showing the Tc values for the two highlighted DELs. d) Density landscapes of the two highlighted DELs, on the left: GTM landscapes, on the right: landscapes predicted by CoLiNN.

one are very similar to each other. These results together clearly show that density landscapes predicted by CoLiNN provide almost the same ranking by similarity to a reference database (here ChEMBL) as true GTM-calculated ones. Such correspondence between rankings confirmed that there is no longer the need to enumerate combinatorial libraries to be able to analyse and compare them with respect to a reference database – CoLiNN allows to do it without.

4.3 | CoLiNN vs GTM: Calculation Time

CoLiNN was trained on a single GPU of type NVIDIA RTX A6000, CUDA version 12.2, GPU memory 48 GB. The physical time needed to train global and local CoLiNN models as well as the physical prediction time expressed in ms/molecule/GPU are given in Table 2 below. It should be noted that the major difference in time between global and local models is due to the fact that the mixed precision option was used for both training and inference of the global model. In more detail, mixed precision training and inference involve using both 16-bit and 32-bit floating-point types for computation, which can lead to increased speed and reduced memory usage.

The enumeration, standardization, descriptor calculation and GTM algorithm launch were performed on a machine powered by dual Intel Xeon Silver 4214 R CPU at 2.40 GHz, total number of CPUs being 48. The typical time needed to perform

TABLE 2 | Physical time spent for training and validation of different CoLiNN models (including calculation of graphs and processing of building block and reaction indices) as well as total physical prediction time using these models.

Model	Total physical training time	Total physical prediction time (ms/molecule/GPU)
Global (388 DELs): 3 M 880 K cmpds	13 h	0.055
Local:1 M cmpds	14 h 37 m	0.124
Local: 50 K cmpds	47 m 35 s	0.091
Local: 25 K cmpds	23 m 17 sec	0.103
Local: 10 K cmpds	21 m 8 sec	0.103

all these calculations expressed in ms/molecule/CPU is given in Table 3. All CoLiNN models provide acceleration from 3000 to 7000-fold compared to traditional enumeration-based workflow. This shows that CoLiNN represents a better option in terms of calculation time/precision ratio for ultra-large combinatorial chemical space visualization when fragment-based descriptors and GTM dimensionality reduction methods are used.

5 | Conclusion

In this work, we present the CoLiNN graph convolutional network capable to predict GTM cumulated responsibilities (chemical space population patterns) for combinatorial libraries, without compound enumeration, thereby enabling the visualization of ultra-large DELs. The „global" CoLiNN model trained on 388 DELs employing 65 K building blocks and 29 reactions was tested on the 2089 DELs that did not participate in training but shared BBs and reactions with the training set libraries. For most of DELs, the predicted projection patterns have high similarity to the ones generated by the GTM algorithm. The rankings of DELs according to their similarity to ChEMBL remained highly consistent when replacing true DEL projection vectors with those predicted using CoLiNN (Spearman's $\rho = 0.956$).

The obtained results demonstrate that CoLiNN holds the potential to become an essential tool for combinatorial compound library design, as it can efficiently explore the library design space without requiring compound enumeration. This can be especially advantageous for DEL technology, where medicinal chemists must select the most promising design from thousands of possibilities. With CoLiNN, different DEL compositions, based on different sets of building blocks/reactions can be quickly assessed, providing instant visualization of their chemical space coverage, in order to identify the best options for DEL design.

TABLE 3 | Physical time spent on each step of the traditional enumeration-based workflow of compound visualization on the GTM.

Task	Physical time (ms/molecule/CPU)
Enumeration using eDesigner*	0.24
Standardization	187.2
Descriptor calculation	207.36
Projection	0.288
Total	395.08

*Runs only on a single process.

These results showcase the high performance of CoLiNN in predicting the combinatorial chemical space motif projected on the GTM without compound enumeration. This tool can be used by medicinal chemists for library design accelerating the process of testing different building block combinations, visualization of ultra-large combinatorial chemical spaces, etc. The developed code can be used in combination with any earlier reported GTM implementation, e.g., the ISIDA/GTM program [33] used in this work or the open-source *ugtm* tool by H. Gaspar [50].

As future development current CoLiNN architecture, the reactions may be represented by Condensed Graphs of Reaction [51], extending the application of CoLiNN to new unseen reactions. Potentially, CoLiNN can be applied to predict projection using other than GTM dimensionality reduction methods; this is an object of future investigations. In addition, prospectively, CoLiNN can be implemented not only for visualization but for the prediction of any property of combinatorial compounds, for example, docking score.

Acknowledgments

The authors are grateful to eMolecules Inc. for the provided library of commercially available BBs, used for DNA-encoded libraries design. We thank Dr. Alexey Orlov for stimulating discussion.

Data Availability Statement

The Python code of CoLiNN is available in the following GitHub repository: https://github.com/Laboratoire-de-Chemoinformatique/CoLiNN The data used in this work are available in the public domain resources: biologically relevant compounds from ChEMBL [52] (version 28) – https://www.ebi.ac.uk/chembl/, collection of eMolecules [38] building blocks partially available on the site https://www.emolecules.com/products/building-blocks, collection of Enamine [39] building blocks available on the site https://enamine.net/building-blocks. The ISIDA/GTM program is available from the authors upon the request.

References

1. A. Neumann, L. Marrison, and R. Klein, "Relevance of the Trillion-Sized Chemical Space "Explore" as a Source for Drug Discovery", *ACS*

Medicinal Chemistry Letters 14, no. 4 (2023): 466–472, https://doi.org/10.1021/acsmedchemlett.3c00021.

2. W. A. Warr, M. C. Nicklaus, C. A. Nicolaou, and M. Rarey, "Exploration of Ultralarge Compound Collections for Drug Discovery", *Journal of Chemical Information and Modeling* 62, no. 9 (2022): 2021–2034, https://doi.org/10.1021/acs.jcim.2c00224.

3. R. M. Franzini, and C. Randolph, "Chemical Space of DNA-Encoded Libraries", *Journal of Medicinal Chemistry* 59, no. 14 (2016): 6629–6644 https://doi.org/10.1021/acs.jmedchem.5b01874

4. Y. Shi, Y. Wu, J. Yu, W. Zhang, and C. Zhuang, "DNA-Encoded Libraries (DELs): A Review of on-DNA Chemistries and their Output", *RSC Advances* 11, no. 4 (2021): 2359–2376, https://doi.org/10.1039/D0RA09889B.

5. N. Favalli, G. Bassi, J. Scheuermann, and D. Neri, "DNA-Encoded Chemical Libraries – Achievements and Remaining Challenges" *FEBS Letters Wiley Blackwell* (2018): 2168–2180, https://doi.org/10.1002/1873-3468.13068.

6. A. L. Satz, "What Do You Get from DNA-Encoded Libraries?," *ACS Medicinal Chemistry Letters* (2018): 408–410, https://doi.org/10.1021/acsmedchemlett.8b00128.

7. R. A. Goodnow, C. E. Dumelin, and A. D. Keefe, "DNA-Encoded Chemistry: Enabling the Deeper Sampling of Chemical Space", *Nature Reviews Drug Discovery* 16, no. 2 (2017): 131–147, https://doi.org/10.1038/nrd.2016.213.

8. L. K. Petersen, A. B. Christensen, J. Andersen, et al., "Screening of DNA-Encoded Small Molecule Libraries inside a Living Cell," *Journal of the American Chemical Society* 143, no. 7 (2021): 2751–2756, https://doi.org/10.1021/jacs.0c09213.

9. R. M. Franzini, D. Neri, and J. Scheuermann, "DNA-Encoded Chemical Libraries: Advancing beyond Conventional Small-Molecule Libraries", *Accounts of Chemical Research* 47 (2014): 1247–1255, DOI: 10.1021/ar400284t

10. M. A. Clark, R. A. Acharya, C. C. Arico-muendel, et al., "Design, Synthesis and Selection of DNA-Encoded Small-Molecule Libraries, 5, no. 9 (2009): 647–655, https://doi.org/10.1038/nchembio.211.

11. A. Gironda-Martínez, E. J. Donckele, F. Samain, and D. Neri, "DNA-Encoded Chemical Libraries: A Comprehensive Review with Succesful Stories and Future Challenges", *ACS Pharmacology & Translational Science* 4, no. 4 (2021): 1265–1279, https://doi.org/10.1021/acsptsci.1c00118.

12. R. A. Lerner, and S. Brenner, "DNA-Encoded Compound Libraries as Open Source: A Powerful Pathway to New Drugs", *Angewandte Chemie International Edition* 56, no. 5 (2017): 1164–1165, https://doi.org/10.1002/anie.201612143.

13. D. Madsen, C. Azevedo, I. Micco, L. K. Petersen, N. Jakob, V. Hansen, "An Overview of DNA-Encoded Libraries: A Versatile Tool for Drug Discovery," 1st ed.; Elsevier B. V., 2020; Vol. 59. https://doi.org/10.1016/bs.pmch.2020.03.001.

14. V. Kunig, M. Potowski, A. Gohla, and A. Brunschweiger, "DNA-Encoded Libraries – An Efficient Small Molecule Discovery Technology for the Biomedical Sciences", *Biol. Chem.* 399, no. 7 (2018): 691–710, DOI: 10.1515/hsz-2018-0119.

15. R. A. Goodnow, *A Handbook for DNA-Encoded Chemistry: Theory and Applications for Exploring Chemical Space and Drug Discovery*, Wiley, Hoboken, New Jersey, 2014, DOI: 10.1002/9781118832738.

16. R. Liu, X. Li, and K. S. Lam, "Combinatorial Chemistry in Drug Discovery", *Current Opinion in Chemical Biology* 38 (2017): 117–126, https://doi.org/10.1016/j.cbpa.2017.03.017.

17. B. Suay-García, J. I. Bueso-Bordils, A. Falcó, G. M. Antón-Fos, and P. A. Alemán-López, "Virtual Combinatorial Chemistry and Pharmacological Screening: A Short Guide to Drug Design", *International Journal of Molecular Sciences. MDPI* (2022): https://doi.org/10.3390/ijms23031620.

18. R. Pikalyova, Y. Zabolotna, D. M. Volochnyuk, D. Horvath, G. Marcou, and A. Varnek, "Exploration of the Chemical Space of DNA-encoded Libraries", *Molecular Informatics* 41, no. 6 (2022): 2100289, https://doi.org/10.1002/minf.202100289.

19. R. Pikalyova, Y. Zabolotna, D. Horvath, G. Marcou, and A. Varnek, "Chemical Library Space: Definition and DNA-Encoded Library Comparison Study Case", *Journal of Chemical Information and Modeling* 63, no. 13 (2023): 4042–4055, https://doi.org/10.1021/acs.jcim.3c00520.

20. R. Pikalyova, Y. Zabolotna, D. Horvath, G. Marcou, and A. Varnek, "Meta-GTM: Visualization and Analysis of the Chemical Library Space", *Journal of Chemical Information and Modeling* 63, no. 17 (2023): 5571–5582, https://doi.org/10.1021/acs.jcim.3c00719.

21. M. Rarey, and J. S. Dixon, "Feature Trees: A New Molecular Similarity Measure Based on Tree Matching", *Journal of Computer Aided Molecular Design* 12, no. 5 (1998): 471–490, https://doi.org/10.1023/A:1008068904628.

22. J. Lübbers, U. Lessel, and M. Rarey, "Enhanced Calculation of Property Distributions in Chemical Fragment Spaces", *Journal of Chemical Information and Modeling* 64, no. 6 (2024): 2008–2020, https://doi.org/10.1021/acs.jcim.4c00147.

23. L. Bellmann, P. Penner, and M. Rarey, "Topological Similarity Search in Large Combinatorial Fragment Spaces", *Journal of Chemical Information and Modeling* 61, no. 1 (2021): 238–251, https://doi.org/10.1021/acs.jcim.0c00850.

24. L. Bellmann, P. Penner, M. Gastreich, and M. Rarey, "Comparison of Combinatorial Fragment Spaces and Its Application to Ultralarge Make-on-Demand Compound Catalogs", *Journal of Chemical Information and Modeling* 62, no. 3 (2022): 553–566, https://doi.org/10.1021/acs.jcim.1c01378.

25. L. Bellmann, R. Klein, and M. Rarey, "Calculating and Optimizing Physicochemical Property Distributions of Large Combinatorial Fragment Spaces", *Journal of Chemical Information and Modeling* 62, no. 11 (2022): 2800–2810, https://doi.org/10.1021/acs.jcim.2c00334.

26. D. K. Agrafiotis, and V. S. Lobanov, "Multidimensional Scaling of Combinatorial Libraries without Explicit Enumeration", *Journal of Computational Chemistry* 22, no. 14 (2001): 1712–1722, https://doi.org/10.1002/jcc.1126.

27. J. F. M. Svensen, *GTM: The Generative Topographic Mapping*, (University of Aston in Birmingham, 1998).

28. H. A. Gaspar, I. I. Baskin, G. Marcou, D. Horvath, and A. Varnek, "GTM-Based QSAR Models and Their Applicability Domains", *Molecular Informatics* 34, no. 6–7 (2015): 348–356, https://doi.org/10.1002/minf.201400153.

29. C. M. Bishop, M. Svensén, and C. K. I. Williams, "GTM: The Generative Topographic Mapping", *Neural Computation* 10, no. 1 (1998): 215–234, https://doi.org/10.1162/089976698300017953.

30. A. A. Orlov, E. V. Khvatov, A. A. Koruchekov, et al., "Getting to Know the Neighbours with GTM: The Case of Antiviral Compounds," *Molecular Informatics* 38, no. 5 (2019): 1–12, https://doi.org/10.1002/minf.201800166.

31. P. Tino, and I. Nabney, "Hierarchical GTM: Constructing Localized Nonlinear Projection Manifolds in a Principled Way", *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, no. 5 (2002): 639–656, https://doi.org/10.1109/34.1000238.

32. H. A. Gaspar, I. I. Baskin, G. Marcou, D. Horvath, and A. Varnek, "Chemical Data Visualization and Analysis with Incremental Generative Topographic Mapping: Big Data Challenge", *Journal of Chemical Information and Modeling* 55, no. 1 (2015): 84–94, https://doi.org/10.1021/ci500575y.

33. D. Horvath, G. Marcou, and A. Varnek, “Generative Topographic Mapping in Drug Design,” *Drug Discovery Today Technologies* xxx, no. xx (2020): 1–9, https://doi.org/10.1016/j.ddtec.2020.06.003.

34. M. van der Laurens, and G. Hinton, “Visualizing Data Using T-SNE”, *Journal of Machine Learning Research* 9 (2008): 2579–2605.

35. D. Miljković, “Brief Review of Self-Organizing Maps,” 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia (2017) 1061–1066, DOI: 10.23919/MIPRO.2017.7973581.

36. D. Digles, and G. F. Ecker, “Self-Organizing Maps for in Silico Screening and Data Visualization”, *Molecular Informatics* 30, no. 10 (2011): 838–846, https://doi.org/10.1002/minf.201100082.

37. A. A. Orlov, T. N. Akhmetshin, D. Horvath, G. Marcou, and A. Varnek, “From High Dimensions to Human Insight: Exploring Dimensionality Reduction for Chemical Space Visualization,” *Molecular Informatics* (2025): 44, e202400265, https://doi.org/10.1002/minf.202400265.

38. eMolecules Inc. https://www.emolecules.com/2020.

39. Enamine Ltd. https://enamine.net/2020.

40. A. Martín, and C. A. Nicolaou, “Navigating the DNA Encoded Libraries Chemical Space,” *Communication Chemistry* 3, (2020): 127, https://doi.org/10.1038/s42004-020-00374-1

41. Y. Zabolotna, A. Lin, D. Horvath, G. Marcou, D. M. Volochnyuk, and A. Varnek, “Chemography: Searching for Hidden Treasures”, *Journal of Chemical Information and Modeling* 61, no. 1 (2021): 179–188, https://doi.org/10.1021/acs.jcim.0c00936.

42. T. Akhmetshin, A. Lin, D. Mazitov, et al., “HyFactor: A Novel Open-Source, Graph-Based Architecture for Chemical Structure Generation,” *Journal of Chemical Information and Modeling* 62, no. 15 (2022): 3524–3534, https://doi.org/10.1021/acs.jcim.2c00744.

43. Indigo Toolkit. https://lifescience.opensource.epam.com/indigo/, Accessed 04. 03. 2024.

44. M. Fey, and J. E. Lenssen, “Fast Graph Representation Learning with PyTorch Geometric. arXiv preprint arXiv:1903.024282019.

45. D. Mendez, A. Gaulton, A. P. Bento, et al., “ChEMBL: Towards Direct Deposition of Bioassay Data,” *Nucleic Acids Research* 47, no. D1 (2019): D930–D940, https://doi.org/10.1093/nar/gky1075.

46. D. Hendrycks, and K. Gimpel, Gaussian Error Linear Units (GELUs) 2016.

47. T. N. Kipf, and M. Welling, Semi-Supervised Classification with Graph Convolutional Networks 2016.

48. M. Lee, “Mathematical Analysis and Performance Evaluation of the GELU Activation Function in Deep Learning”, *Journal of Mathematics* 2023, no. 1 (2023): 4229924, https://doi.org/10.1155/2023/4229924.

49. A. H. Huang, Expanded Gating Ranges Improve Activation Functions. other.

50. H. A. Gaspar, Ugtm: A Python Package for Data Modeling and Visualization Using Generative Topographic Mapping. 2018.

51. F. Hoonakker, N. Lachiche, A. Varnek, and A. Wagner, “Condensed Graph of Reaction: Considering a Chemical Reaction as One Single Pseudo Molecule”, *International Journal on Artificial Intelligence Tools* 20, no. 2 (2011): 253–270, https://doi.org/10.1142/S0218213011000140.

52. A. Gaulton, L. J. Bellis, A. P. Bento, et al., “ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery,” *Nucleic Acids Research* 40, no. D1 (2012): D1100–D1107, https://doi.org/10.1093/nar/gkr777.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.